Comprehensive Insights in the *Mycobacterium avium* subsp. *paratuberculosis* Genome Using New WGS Data of Sheep Strain JIII-386 from Germany

Petra Möbius[1,*,†], Martin Hölzer[2,†], Marius Felder[3], Gabriele Nordsiek[4], Marco Groth[3], Heike Köhler[1], Kathrin Reichwald[3], Matthias Platzer[3], and Manja Marz[2,3,5]

[1]NRL for Paratuberculosis, Institute of Molecular Pathogenesis, Friedrich-Loeffler-Institut (Federal Research Institute for Animal Health), Jena, Germany

[2]RNA Bioinformatics and High Throughput Analysis, Faculty of Mathematics and Computer Science, Friedrich Schiller University Jena, Germany

[3]Leibniz Institute for Age Research – Fritz-Lipmann-Institute (FLI), Jena, Germany

[4]Department of Genome Analysis, Helmholtz Centre for Infection Research, Braunschweig, Germany

[5]Michael Stifel Center, Jena, Germany

*Corresponding author: E-mail: petra.moebius@fli.bund.de.

†These authors contributed equally to this work.

Accepted: August 5, 2015

Abstract

Mycobacterium avium (*M. a.*) subsp. *paratuberculosis* (MAP)—the etiologic agent of Johne's disease—affects cattle, sheep, and other ruminants worldwide. To decipher phenotypic differences among sheep and cattle strains (belonging to MAP-S [Type-I/III], respectively, MAP-C [Type-II]), comparative genome analysis needs data from diverse isolates originating from different geographic regions of the world. This study presents the so far best assembled genome of a MAP-S-strain: Sheep isolate JIII-386 from Germany. One newly sequenced cattle isolate (JII-1961, Germany), four published MAP strains of MAP-C and MAP-S from the United States and Australia, and *M. a.* subsp. *hominissuis* (MAH) strain 104 were used for assembly improvement and comparisons. All genomes were annotated by BacProt and results compared with NCBI (National Center for Biotechnology Information) annotation. Corresponding protein-coding sequences (CDSs) were detected, but also CDSs that were exclusively determined by either NCBI or BacProt. A new Shine–Dalgarno sequence motif (5′-AGCTGG-3′) was extracted. Novel CDSs including PE-PGRS family protein genes and about 80 noncoding RNAs exhibiting high sequence conservation are presented. Previously found genetic differences between MAP-types are partially revised. Four of ten assumed MAP-S-specific large sequence polymorphism regions (LSP^{S}s) are still present in MAP-C strains; new LSP^{S}s were identified. Independently of the regional origin of the strains, the number of individual CDSs and single nucleotide variants confirms the strong similarity of MAP-C strains and shows higher diversity among MAP-S strains. This study gives ambiguous results regarding the hypothesis that MAP-S is the evolutionary intermediate between MAH and MAP-C, but it clearly shows a higher similarity of MAP to MAH than to *Mycobacterium intracellulare*.

Key words: MAP-S, Johne's disease, ncRNA, Shine–Dalgarno sequence, new LSP^{S}s, SNV/SNP, evolution of MAP-types.

Introduction

Mycobacteria—a genus of Actinobacteria—include pathogens known to cause serious diseases in man and other mammals: Tuberculosis (*Mycobacterium tuberculosis*) and leprosy (*Mycobacterium leprae*). *Mycobacterium tuberculosis* has been in the focus of research for a long time, aiming to fight against tuberculosis worldwide. Therefore, most new scientific findings concerning Mycobacteria are based on this species. Another *Mycobacterium* species which is distributed worldwide affects domestic and wild ruminants: *Mycobacterium avium* subsp. *paratuberculosis* (MAP).

MAP is the causative agent of paratuberculosis (Johne's disease); a chronic granulomatous enteritis causing malnutrition, therapy-resistant diarrhea, emaciation, low milk yield, and ultimately death (Clarke and Little 1996).

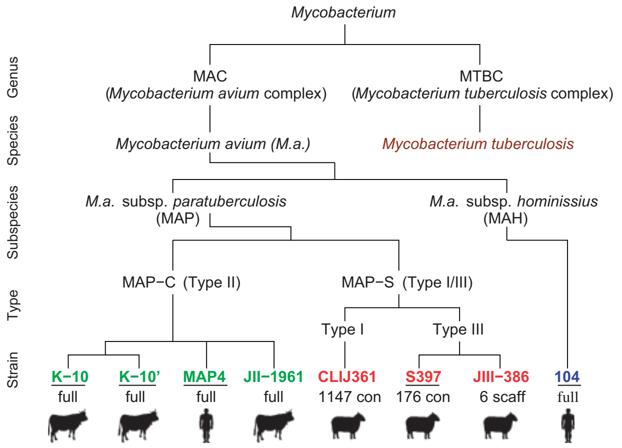

Fig. 1.—Overview of *M. avium* strains compared in this study and their assembly and annotation status. Strains of MAP-S (Type I/III): JIII-386, S397, CLIJ361 (red); strains of MAP-C (Type II): K-10, K-10′, MAP4, JII-1961 (green); MAH strain 104 (blue); *M. tuberculosis* strain H37Rv (brown, used for extended ncRNA annotation). Underlined, annotations available; scaff, scaffolds; con, contigs; full, finished genome. Pictograms describe host origin.

Paratuberculosis is of considerable economic significance especially for the dairy industry.

MAP belongs to the species *M. avium* together with the bird pathogens *M. avium* subsp. *avium* (MAA) and *M. avium* subsp. *silvaticum* (MAS), as well as the related environmental organism *M. avium* subsp. *hominissuis* (MAH) associated with opportunistic infections in man and pig (Thorel et al. 1990; Mijs et al. 2002). MAH shows the highest genetic variability within this species (Turenne et al. 2008). Comparing strains of different *M. avium* subspecies a high diversity can be recognized (Wu et al. 2006; Paustian et al. 2008; Hsu et al. 2011). In contrast, strains of subspecies *paratuberculosis* exhibit a relatively low genetic heterogeneity. Depending on the study, MAP was differentiated into two or three groups based on phenotype, genotype, and host association: MAP Type I + III and MAP Type II (see fig. 1), also designated as MAP-S (sheep type) and MAP-C (cattle type), respectively (Collins et al. 1990; Stevenson et al. 2002; Turenne et al. 2008; Castellanos et al. 2009).

MAP Type I strains predominantly infect ovine hosts; MAP Type II strains principally infect cattle but also deer, goat, sheep, and other ruminants (Whittington et al. 2000; Sevilla et al. 2007; Fritsch et al. 2012). MAP Type III strains (intermediate) are closely related to Type I strains and have been isolated up to now from sheep, goat, cattle, and camels (De Juan et al. 2005, 2006; Möbius et al. 2009; Castellanos et al. 2010; Ghosh et al. 2012).

The high entry of the MAP organism into the immediate environment of diseased animals by shedding and the high tenacity of MAP within the environment (Whittington et al. 2004) generate an increasing risk of exposure to MAP not only for ruminants but also for other mammals. For instance, MAP was detected in tap water, rivers, dams (Pickup et al. 2006; Rhodes et al. 2014), and also in raw milk (Shankar et al. 2010). Furthermore, MAP was isolated from a clinically diseased donkey (Stief et al. 2012). MAP has also been isolated from man. Its etiologic role in Crohn's disease is under discussion (Over et al. 2011; Atreya et al. 2014).

About 10 years ago, first sequence data of MAP strains were published. Isolate MAP K-10 from the United States—belonging to the MAP-C group—was fully sequenced (Li L, Bannantine JP, et al. 2005). Later on it was resequenced and better annotated by Wynne et al. (2010). Recently, ovine-derived isolates CLIJ361 (MAP-S, Type I) from Australia (Wynne et al. 2011), MAP S397 (MAP-S, Type III) from the United States (Bannantine et al. 2012), and the first human-derived isolate MAP4 (MAP-C) from the United States (Bannantine et al. 2014) have been sequenced.

The availability of these MAP-S genome sequences—although not fully assembled—improved the informational value of genome comparisons no longer only based on MAP K-10 and *M. avium* strain 104. In the meantime, MAP-S- and MAP-C-specific loci, genome deletions and insertions have been identified and evolutionary relationships proposed (Dohmann et al. 2003; Marsh et al. 2006; Semret et al. 2006; Paustian et al. 2008; Alexander et al. 2009; Bannantine et al. 2012).

Besides the comparative whole-genome sequence (WGS) analysis, in the past decade non-protein-coding fractions of the transcriptome were studied in bacteria (Sharma et al. 2010; Lechner et al. 2014; Wehner, Mannala, et al. 2014). Regarding Mycobacteria, noncoding RNAs (ncRNAs) were identified in *M. tuberculosis* and their role in the regulation of the pathogen metabolism was studied (Arnvig and Young 2009, 2010). Furthermore, RNA sequences were analyzed in *M. avium* including MAH and MAA (Ignatov et al. 2013). Until now, no data have been published on the full set of ncRNAs in MAP.

The objective of this study was to sequence a further MAP-S strain: The ovine-derived strain JIII-386 from Germany (Europe), and to compare sequence data with seven assemblies of related genomes from other continents to examine previously defined genomic differences between MAP-S and MAP-C strains (Type I/III and II, respectively). Complete genome sequences of a bovine-derived MAP-C strain (also from Germany) and the *M. avium* strain 104 were included. A genome-wide annotation of protein-coding sequences (CDSs) was performed by using two data resources, NCBI (National Center for Biotechnology Information) and BacProt. For the first time, a comprehensive annotation of regulatory RNAs in MAP was performed. Based on the current data analysis, we wanted to find out new aspects regarding proposed ancestral relationship of *M. avium* complex (MAC) strains and indications for an evolution or conservation of regulatory RNAs.

Materials and Methods

MAP Ovine Isolate JIII-386 and Bovine Isolate JII-1961

Isolation, identification, and characterization of the MAP-S, Type III isolate JIII-386 were described by Möbius et al. (2009). The strain was isolated in 2003 and belongs to the strain collection of the Friedrich-Loeffler-Institut in Jena (Germany). Briefly, JIII-386 was isolated from ileal mucosa of a sheep from a migrating herd in the north-west of Germany. The animal showed no clinical symptoms. Paratuberculosis was suspected based on positive serological results, detection of MAP in feces by culture, and pathomorphological and histological results after necropsy. JIII-386 had been cultured using modified Middlebrook 7H11 solid medium (Difco) containing 10% oleic acid-albumin-dextrose-catalase (OADC), Amphotericin B, and Mycobactin J (Allied Monitor, Fayette, NY). Subcultivation was done on modified Loewenstein–Jensen solid medium, also supplemented with Mycobactin J.

Additionally, MAP strain JII-1961 (MAP-C) isolated from cattle in 2003 and sequenced and assembled on the chromosomal level at the Helmholtz Centre for Infection Research (Braunschweig, Germany) was included and annotated in this study. This isolate originated from the ileocecal lymph node of a clinically diseased dairy cow from a paratuberculosis positive herd in eastern Germany. JII-1961 was isolated and subcultivated using Herrold's Egg Yolk Medium (HEYM) supplemented with Mycobactin J.

JIII-386 had been grown for up to 7 months, strain JII-1961 for up to 6 weeks. Both isolates were characterized by positive acid-fast staining and their growth characteristics and were proved to be MAP by cultural confirmation of mycobactin-dependency and detection of the presence of the IS*900* insertion sequence using polymerase chain reaction (PCR) (Englund et al. 2001).

The genotypes of isolates were determined (Möbius et al. 2009) by multitarget genotyping based on IS*900*-RFLP (restriction fragment length polymorphism) (four digestion enzymes), mycobacterial interspersed repetitive unit-variable-number of tandem-repeat (MIRU-VNTR) (nine loci), and simple sequence repeat (four loci) analysis (Amonsin et al. 2004; Thibault et al. 2007; Möbius et al. 2008). Isolates were expanded for sequencing on HEYM supplemented with Mycobactin J. Genomic DNA was prepared by the cetyltrimethylammonium bromide method described by van Soolingen et al. (1991), and identity of the strain was confirmed by MIRU-VNTR-genotyping.

Sequencing

Whole-genome shotgun sequencing was performed. Illumina paired-end (fragment size ~300 bp) and mate-pair (fragment size ~2.2 kb) libraries were generated from fragmented genomic DNA of MAP strain JIII-386. Libraries were sequenced using Illumina GAIIx (paired-end library) and HiSeq2000 (mate-pair library) and resulted in 28.6 million 101-bp paired-ends (~1.100-fold genome coverage) and 10.9 million 100-bp mate-pairs (~440-fold genome coverage) (see supplementary table S3, Supplementary Material online (http://www.rna.uni-jena.de/supplements/mycobacterium/, last accessed August 21, 2015).

Data Preprocessing and De Novo Assembly

After quality trimming and removal of duplicons, 2 × 27.5 million paired-end reads and 2 × 10.5 million mate-pairs were de novo assembled using CLC Genomics Workbench (v5.0, default parameter; http://www.clcbio.com, last accessed August 21, 2015). This initial assembly (I) resulted in 130 contigs with a total length of 4,792,650 bp. The assembly was improved with the scaffolding tool SSPACE v2.0 (Boetzer et al. 2011) using all mate-pairs. To close remaining sequence gaps, primers flanking missing regions were designed. The amplicons obtained from genomic DNA were sequenced directly using Sanger technology. The resulting assembly (II) comprises 14 scaffolds totaling in 4,846,897 bp (see supplementary table S4, Supplementary Material online).

Assembly Improvement

To improve the assembly (II), the following steps to handle low-coverage regions, low-quality reads, misassemblies, replacing gap regions, and connecting scaffolds were applied: Four additional de novo genome assembly tools were separately used on both libraries: Velvet (v1.2.10, $k = 55$) (Zerbino and Birney 2008), ABySS (v1.3.4, $k = 45$) (Simpson et al. 2009), SPAdes mainly implemented for single-cell data (v2.5.1, $k = 43{,}55{,}65$) (Bankevich et al. 2012)—all de Bruijn graph based (Li et al. 2012)—and the seed-and-extend approach based JR-Assembler (v1.0.3, default parameters) (Chu et al. 2013). The resulting contigs (>1,000 bp) were merged and clustered for sequence similarities using cd-hit-est (v4.6, -c 0.95) (Li and Godzik 2006) to reduce redundancy. Statistical information and each assembly can be retrieved from the supplementary table S4, Supplementary Material online.

Related Genomes

Related genomes served as reference genomes in this study to assist in assembly, open-reading frame (ORF) predictions, and annotation. Furthermore, they were used for comparison of different MAP types including strains originating from different geographic regions of the world. The selection comprises the genomic sequences of the three MAP-C (Type II) strains: K-10/K-10′ (Li L, Bannantine JP, et al. 2005; Wynne et al. 2010), MAP4 (Bannantine et al. 2014) and JII-1961 (Möbius P, Jarek M, Köhler H, Nordsiek G, unpublished data), two sheep-derived MAP-S strains, one of Type I: CLIJ361 (Wynne et al. 2011) and one of Type III: S397 (Bannantine et al. 2012), as well as one MAH strain: *M. avium* strain 104 designated as MAH 104 (Yakrus and Good 1990). Strains K-10, MAP4, and S397 originated from the United States, strain CLIJ361 from

Australia, and strain JII-1961 from Germany. Genotypes of investigated MAP isolates within this study are shown in supplementary table S2, Supplementary Material online. Currently, finished genome sequences of these three MAP-C strains are available. The two ovine isolates are available at contig level: S397 which comprises 176 contigs and CLIJ361 based on 1,147 contigs (draft genomes).

All reference sequences and available annotation files for the above-mentioned strains (except JII-1961) were downloaded from NCBI. All used genome data are linked in the supplementary table S1, Supplementary Material online.

Mauve (v2.3.1) (Darling et al. 2004, 2010) was used for genome alignments of strains K-10′, JIII-386, and S397 and comparison of examined MAP strains.

Annotation of CDSs

Annotations for MAP strains K-10, K-10′, MAP4, S397 and strain MAH 104 were downloaded from NCBI (see supplementary table S1, Supplementary Material online). For reference-based annotation of CDSs, BacProt (unpublished data) based on Proteinortho (Lechner 2009; Lechner et al. 2011) was used to complement present annotations. Furthermore, the novel ORF prediction of BacProt, containing Shine–Dalgarno and Pribnow box motif information, was applied. For each *M. avium* strain, reannotated and previously annotated ORFs as well as statistics like codon usage and occurrence of Shine–Dalgarno sequence motifs were calculated. For the ovine-derived strain JIII-386, annotation was complemented with data from Bannantine et al. (2012) by using BLAST (Basic Local Alignment Search Tool) (Altschul et al. 1990) (v2.2.27+, E value $\leq 10^{-4}$) with at least 90% identity and an alignment length of 90%.

ORFs with sequence homology to genes with an assigned function in the NCBI annotation were identified and designated as CDSs.

For each isolate, annotations provided by NCBI were merged with the BacProt annotations to find ORFs being present or absent between two strains by using BLAST (E value $\leq 10^{-4}$). All ORFs of strain $\mathcal{A}$, which could not achieve a sequence overlap of at least 50% in length and identity against the genome of strain $\mathcal{B}$, were marked as present in $\mathcal{A}$ but absent in $\mathcal{B}$. ORFs without an assigned function were excluded from supplementary table S15*a*, Supplementary Material online. These data provide an overview of the different ORFs, present/absent between the investigated *M. avium* strains. Detailed analyses of single genes and gene clusters as well as large sequence polymorphisms (LSPs) and phylogenetic relationships were performed by more restrictive parameters (E value $\leq 10^{-20}$, alignment length $\geq$ 95% of query, sequence similarity $\geq$ 90%; depending on the kind of analysis) and manual investigation of all BLAST results, alignments, and sequences.

Single nucleotide variants (SNVs) were searched by pairwise comparison of CDSs of the eight investigated genomes. First, BLAST (E value $\leq 10^{-4}$) was used to assign homologous sequences between two strains, which were aligned in a second step using MAFFT (v.7.017b, method: L-INS-i) (Katoh et al. 2002). The resulting alignments were searched for SNVs by individual ruby scripts.

The presence or absence of 35 LSPs, each containing several ORFs and previously reported by Alexander et al. (2009) and Bannantine et al. (2012), was examined by using BLASTn+ across the investigated strains.

Annotation of ncRNAs

ncRNAs were annotated by homology search of Rfam (v.11.0) (Gardner et al. 2009) families using the Genomewide RNA Annotation Pipeline (GORAP) (unpublished data), which currently comprises Infernal (v1.1) (Nawrocki et al. 2009), Bcheck (v0.6) (Yusuf et al. 2010), RNAmmer (v1.2) (Lagesen et al. 2007), and tRNAscan-SE (v1.3.1) (Lowe and Eddy 1997) for detection of different ncRNA classes. Within the pipeline, family-specific parameters and several filter steps based on taxonomy, secondary structure, and primary sequence comparison were used. To compare the amount of ncRNAs, GORAP was used to perform additional annotation of ncRNAs for two well-known bacteria: *Escherichia coli* and *Streptococcus entericus*. All resulting Stockholm alignments were hand-curated with the help of Emacs RALEE mode (Griffiths-Jones 2005).

Phylogenetic Reconstruction/Ancestral Relationship

To obtain the relationship between all investigated *M. avium* strains, a phylogenetic reconstruction based on a selected set of CDSs and ncRNAs shared by all strains was performed. For CDSs, BLASTn+ (E value $\leq 10^{-10}$, alignment length >90% of query) and the extended annotations by BacProt were used to find coding sequences that are common between all species. From this a set of 790 CDSs was obtained, which was aligned on nucleotide (~930,000 nt per species) and amino acid (~310,000 amino acids) level using MAFFT. Furthermore, the ncRNAs annotated by GORAP were used to align a set of 70 ncRNAs (~8,200 nt) with MAFFT (L-INS-i). Maximum-likelihood tree constructions were performed on all three alignments using RAxML (v8.0.25) (Stamatakis 2014) with the GTRGAMMA model for nucleotide alignments and PROTGAMMAWAG for amino acids. All calculations were applied with 1,000 bootstrap replicates and outgroup rooting (*M. tuberculosis* H37Rv). The Newick Utilities suite (v1.6) (Junier and Zdobnov 2010) was used to visualize the calculated trees.

Results and Discussion

Genome Sequencing, Assembly, and Analysis

The general genomic features of MAP JIII-386 are presented in table 1 together with the data for the other investigated strains.

Table 1
General Genome Features of the Different *Mycobacteria* Strains

Type	MAP-C				MAP-S			MAH
Strain	K-10	K-10’	MAP4	JII-1961	JIII-386	S397	CLIJ361	104
Origin	🐄	🐄	🧍	🐄	🐑	🐑	🐑	🧍
General Features								
Genome size (bp)	4 829 781	4 832 589	4 829 424	4 829 628	4 850 274	4 813 711	4 612 386	5 475 491
Assembly status	1 Chr	1 Chr	1 Chr	1 Chr	6 Scaff	176 Con	1147 Con	1 Chr
N50	n.a.	n.a.	n.a.	n.a.	1245802	56150	7088	n.a.
Max con/scaff	4 829 781	4 832 589	4 829 424	4 829 628	1 505 968	137 410	49 981	5 475 491
G+C-content (%)	69.3	69.3	69.3	69.3	69.16	69.31	68.96	68.99
Protein-coding ORF annotation by `BacProt`								
Homologous ORFs	3096	3081	3082	3099	3067	3008	2458	3553
Hypothetical ORFs	952	967	960	948	991	3569	5547	1054
Housekeeping ncRNA annotation								
tRNAs	46	46	46	46	46	44	46	46
5S rRNA	1	1	1	1	1	1	1	1
SSU rRNA bacteria	1	1	1	1	1	1	1	1
LSU rRNA bacteria	1	1	1	1	1	1	1	1
RNase P	1	1	1	1	1	1	1	1
tmRNA	1	1	1	1	1	1	1	1
Bacteria small SRP	1	1	1	1	1	1	1	1
other ncRNAs								
PyrR	1	1	1	1	1	1	1	1
6C	1	1	1	1	1	1	0	1
Actino-pnp	1	1	1	1	1	1	1	1
mraW	2	2	2	2	2	2	2	2
ASdes	3	3	3	3	3	3	3	3
ASpks	4	4	4	4	3	3	3	4
F6	1	1	1	1	1	1	1	1
G2	0	0	0	0	1?	1?	1?	1?
AS1890	1?	1?	1?	1?	1?	1?	1?	1?
Riboswitches								
TPP	2	2	2	2	2	2	1	2
Cobalamin	2	2	2	2	2	2	2	2
Glycine	1	1	1	1	1	1	1	1
SAM-IV	1	1	1	1	1	1	1	1
SAH	1	1	1	1	1	1	1	2
pan	1	1	1	1	1	1	1	1
pfl	1	1	1	1	1	1	1	0
ydaO-yuaA	1	1	1	1	1	1	1	1
ykoK	3	3	3	3	3	3	3	3
ykkC-yxkD	1	1	1	1	1	1	1	1
ykkC-III	1	1	1	1	2	2	2	2

Note.—The number of ORFs with a homologous sequence in NCBI (homologous ORFs) and additionally hypothetical ORFs, both predicted by BacProt, are provided. NcRNAs and riboswitches were annotated by homology search of Rfam (v.11.0) (Gardner et al. 2009) families using the GORAP pipeline (unpublished data), see Materials and Methods. For further information (fasta, gff, stk files), see supplementary tables S11, S14, S20, and S22, Supplementary Material online. chr, chromosome; scaff, scaffolds; con, contigs; N50, length of the shortest con/scaff, so that at least 50% of all bp in the assembly are represented by this and all longer contigs; ?, candidate, further analysis needed. TPP binds thiamin pyrophosphate (TPP) to regulate thiamin biosynthesis and transport (Winkler et al. 2002); Cobalamin binds adenosylcobalamin to regulate vitamin B_{12} (cobalamin) biosynthesis and transport (Nahvi et al. 2002); Glycine binds glycine to regulate glycine metabolism genes, including use of glycine as energy source (Mandal et al. 2004); SAM-IV binds *S*-adenosyl methionine (SAM) to regulate methionine as well as SAM biosynthesis/transport (Weinberg et al. 2007); SAH recycling of *S*-adenosylhomocysteine (SAH), produced during SAM-dependent methylation reactions (Weinberg et al. 2007); pan predicted riboswitch function, located in 5′-UTRs of genes encoding enzymes involved in vitamin pantothenate synthesis (Weinberg et al. 2010); pfl predicted riboswitch function, consistently present in genomic locations corresponding to 5′-UTRs of protein-coding genes (Weinberg et al. 2010); ydaO–yuaA, genetic “off” switch for ydaO and yuaA genes, maybe triggered during osmotic shock (Barrick et al. 2004); ykok, MG^{2+}-sensing riboswitch, controls expression of magnesium ion transport proteins (Barrick et al. 2004); ykkC–yxkD, upstream of ykkC and yxkD genes in *Bacillus subtilis* and related genes in other bacteria, function mostly unclear (Weinberg et al. 2010); ykkC-III predicted riboswitch function, appears to regulate genes related to preceding motifs such as ykkC and yxkD (Weinberg et al. 2010); NA, not applicable.

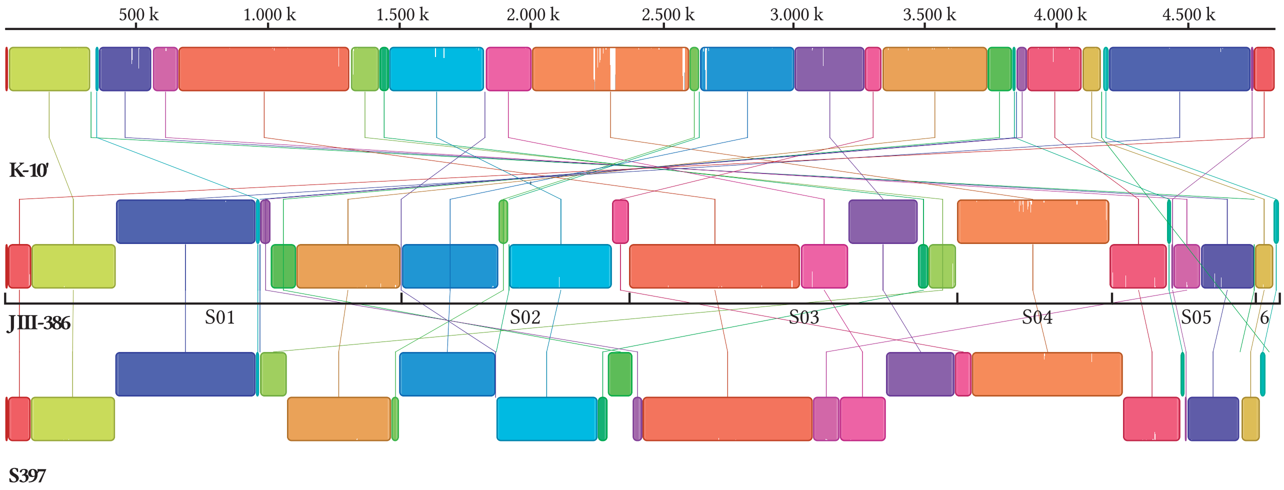

Fig. 2.—Genome comparison of K-10′ (top), JIII-386 (middle), and S397 (bottom) calculated with Mauve. Colored blocks connected by lines indicate homologous regions which are internally free from genomic rearrangements. White areas within blocks indicate sequence regions of lower similarity. Blocks below the center line are aligned reverse complementary. See detailed supplementary figure S10a, Supplementary Material online.

Using the described approach of cluster assembly (see Materials and Methods section) and several genome comparisons with related strains, the assembly (II) of JIII-386 was improved. Thirteen sequences of low complexity, comprising nine poly-N, three poly-A and one poly-T region, were substituted resulting in an overall replacement of 10,254 bp. Five extensions of 5′- and 3′-endings with a minimum of 44 bp and a maximum of 142 bp, respectively, were performed.

The final assembly (III) of JIII-386 comprises 4,850,274 bp on six scaffolds (see also supplementary table S5, Supplementary Material online). Compared with those of the other ovine isolates, it includes fewer gaps and a much better N50 value of 1,245,805 bp. JIII-386 shows a slightly lower G+C content of 69.16% in comparison to those of MAP Type II isolates and also MAP S397 (table 1). Based on the scaffolds of JIII-386, possible connections between contigs within the assemblies of MAP S397 and CLIJ361 could be determined (see supplementary table S8) which might be helpful for further improvement of these assemblies. For calculations and details, see supplementary tables S6–S9, Supplementary Material online.

A graphical visualization of a genome-wide alignment of the K-10′, JIII-386, and S397 sequences is shown in figure 2 and in more detail in supplementary figure S10a, Supplementary Material online. The genomic arrangement of strain JIII-386 is similar to strain S397 (both MAP Type III) and has still large genome fractions in homology with K-10′. Scaffold S03 of strain JIII-386 comprises the longest genomic region (650,087 bp, 558 CDSs annotated with BacProt), which is homologous between the three strains but inversely oriented in JIII-386 and S397 compared with that in K-10′. Two other large homologous regions are located on S01 and S04 of JIII-386 and S397 but in different order than in the K-10′ genome. Two regions greater than 17,000 bp (5′ in S01 and 3′ in S04) were detected in JIII-386 and S397 but not in K-10′ by Mauve alignment (see supplementary fig. S10b, Supplementary Material online). Further analysis showed that these two regions comprise 16 and 15 CDSs, respectively, which are really absent in the investigated MAP-C genomes but present in MAP-S and also in MAH 104 (see supplementary table S17, Supplementary Material online).

Annotation of Protein-Coding Genes

Annotation was performed in two independent steps: 1) BLAST-based lift-over from CDS annotation of MAP S397 (Bannantine et al. 2012) to JIII-386 and 2) semi-de novo annotation through BacProt.

For JIII-386, 4,598 ORFs were predicted using the lift-over annotation based on 4,619 ORFs of strain S397 from NCBI (JIII-386 annotation, see supplementary table S12, Supplementary Material online) and 4,058 ORFs using BacProt (table 2). For JIII-386, the number of genes with assigned function (3,067 CDSs) and those without assigned function (991 "hypothetical ORFs") corresponds to the ORF numbers of all MAP-C isolates and approximately to the ORF numbers of MAH 104 (tables 1 and 2). For MAP S397 and CLIJ361 a higher number of hypothetical ORFs were generated by BacProt, whereby a large amount of these ORFs were detected only on short sequences (~120 bp). This was presumably caused by the BacProt prediction analysis for these two strains based on the published assemblies with a high number of contigs (176 for S397 and >1,000 for CLIJ361).

For all strains except MAP S397 BacProt identified fewer ORFs than provided to date in the NCBI annotations; however, additional ORFs were found (table 2). Both annotations were

Table 2

Annotations Obtained from NCBI and Those Additionally Calculated Using BacProt Lead to an Extended Annotation for Each Investigated *Mycobacterium avium* (Last Column)

subsp.	host	strain	**NCBI**	**BacProt**	Corresponding	Start shifted	End shifted	NCBI only	BacProt only	**Extended**
MAP-C	🐄	K-10	4350	4048	2332	411	458	1149	847	5197
			1146	**3096**	**998**	**60**	**77**	**11**	**1961**	**3107**
	🐄	K-10'	4394	4048	2374	432	433	1155	875	5269
			3048	**3081**	**2046**	**297**	**319**	**385**	**480**	**3527**
	🧍	MAP4	4326	4042	2467	342	382	1135	851	5177
			3029	**3082**	**2179**	**248**	**293**	**309**	**362**	**3391**
	🐄	JII-1961[a]		4047						
				3099						
MAP-S	🐑	JIII-386[b]	4598	4058	2259	479	484	1376	882	5480
			3166	**3067**	**1953**	**363**	**349**	**501**	**428**	**3594**
	🐑	S397	4619	6577	2339	386	458	1436	3394	8013
			3179	**3008**	**2033**	**262**	**326**	**558**	**387**	**3566**
	🐑	CLIJ361[a]		8005						
				2458						
MAH	🧍	104	5120	4607	2846	357	509	1408	895	6015
			3472	**3553**	**2661**	**248**	**373**	**190**	**271**	**3743**

Note.—In the second lines (bold): only predicted ORFs with homology to genes with an assigned function in the NCBI annotation are shown (CDSs). Corresponding, ORFs identified by BacProt and NCBI originating from same positions in the genome; Start/End shifted, ORFs identified by BacProt and NCBI but with differences in length (only 5′ or 3′); NCBI/BacProt only, ORFs identified only by NCBI/BacProt; Extended, total number of ORFs (combination of NCBI + BacProt only). All *.gff files are provided in the supplementary tables S1, S11, and S12, Supplementary Material online.

[a]MAP strains with currently no NCBI annotation available, instead only BacProt results are shown.

[b]Lift-over annotation based on NCBI, MAP S397.

merged to generate an extended prediction of ORFs (table 2, last column). A large fraction of ORFs without an assigned function was included in both annotations; therefore a second line was added to table 2 for each strain in which only the number of ORFs with an assigned function was presented (CDSs). Approximately 49% (first line) and approximately 62% (second line) of all genes (extended panel) were annotated on the same positions by BacProt and NCBI (corresponding genes), whereas approximately 20% (first line) and approximately 9% (second line) of all were found on unique positions with either BacProt or NCBI only. The intersection of corresponding ORFs seems to be more reliable, although the additionally detected ORFs using BacProt were of special interest for further analyses.

For all ORFs annotated by BacProt, the Shine–Dalgarno sequence motif was extracted (see fig. 3), which represents a part of the ribosomal binding site on prokaryotic mRNA. The motif is generally located upstream of a start codon and involved in the recognition of translation start sites during the initial phase of protein synthesis. Remarkably, the Shine–Dalgarno motif observed in all *M. avium* strains examined in this study and additionally in *M. tuberculosis* strain H37Rv: 5′-AGCTGG-3′ (fig. 3) was different from the standard 5′-AGGAGG-3′ pattern (Shine and Dalgarno 1974), possibly conserved for the genus *Mycobacterium*. Using this newly identified Shine–Dalgarno sequence and Pribnow box, ORFs with and without known function were predicted by BacProt and listed in supplementary table S11, Supplementary Material online (see gff-files).

Additionally to the annotation of CDSs an overview of codon usage for each investigated Mycobacteria strain is given in supplementary table S13, Supplementary Material online. As expected, similarities regarding the ratio of G+C-rich codons were found. The codon preferences for G+C correspond likely with the high (almost 70%) G+C content (see table 1) of *M. avium* genomes.

Several regions that are present multiple times in the genome of JIII-386 were discovered. Among these there are insertion sequences (IS), previously described to act as transposable elements, also responsible for the genomic diversity of Mycobacteria (Dale et al. 1995; Sreevatsan et al. 1997; Rindi and Garzelli 2014) and used as molecular epidemiological markers. Seventeen copies of MAP-specific IS*900* (Green et al. 1989; Semret, Turenne, and Behr 2006) were verified in JIII-386, as in strains K-10 and S397 (Bannantine et al. 2012). Additionally, six copies of IS*Map02* were present in JIII-386 as described before for K-10 and S397 by Bannantine et al. (2012). Two copies of IS*Mpa1* (not three copies as detected by Olsen et al. [2004] for MAP strains) were found in the genome of JIII-386. Only five copies of IS*1311* are present in JIII-386, instead of seven copies as reported previously for K-10 and S397 (Bannantine et al. 2012). Furthermore, we found eight copies of IS*1311* in the genomes of K-10′ and MAP4.

Annotation of ncRNAs

In the last decade, ncRNAs, possible regulators of cellular processes, and virulence control (Arnvig and Young 2010; Papenfort and Vogel 2010) gained more importance. They were characterized for *M. tuberculosis* by Arnvig and Young (2009) and Miotto et al. (2012). For *M. avium* (including MAH and MAA), two riboswitches as well as several antisense and intergenic transcripts have been identified (Ignatov et al. 2013).

Hits for all known ncRNAs provided by the Rfam database, based on a screening of the seven MAP genomes and MAH 104, are presented in table 1. All corresponding files are available in stk, gff and fa-format in the supplementary table S19, Supplementary Material online.

In general, ncRNAs among the investigated *M. avium* lineages are extremely conserved (e.g., tRNAs and riboswitches; table 1)—there are only few exceptions such as ASpks and ykkC-III which differ in the number of detected copies

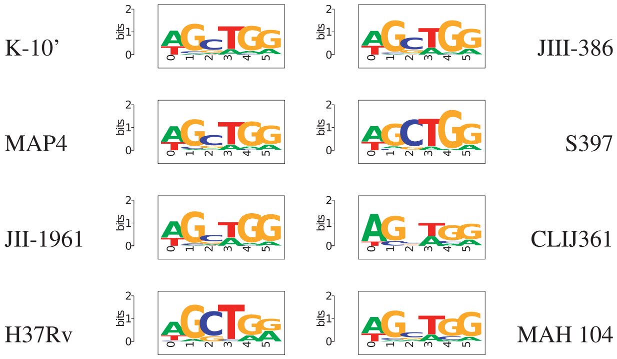

Fig. 3.—Shine–Dalgarno sequence motifs of investigated *M. avium* strains and *M. tuberculosis* strain H37Rv. Detailed information and the motif of MAP strain K-10 (similar to K-10') can be found in supplementary table S11, Supplementary Material online.

between MAP-C, MAP-S, and MAH 104 and are discussed in detail below. The high conservation of ncRNAs between the different *M. avium* strains is remarkable—even other bacteria, like the closely related strains of the obligate intracellular family *Chlamydiaceae* (Sachse et al. 2014), show more differences in their small ncRNA repertoire compared with the *M. avium* strains presented here.

Housekeeping ncRNAs

All examined genomes in this study contained 46 tRNAs in contrast to 45 tRNAs detected by Li et al. (2005), among them three copies of two different methionine- and one selenocysteine-tRNA genes (see supplementary table S21, Supplementary Material online). In contrast to the study of Bannantine et al. (2012) in the genome of strain S397, the genes for Arg-TCT and Lys-TTT could not be identified by tRNAscan-SE analysis (see supplementary table S21, Supplementary Material online). Each of the other housekeeping ncRNAs (ribosomal RNAs [rRNAs], RNase P RNA, transfer-messenger RNA [tmRNA], and signal recognition particle [SRP] RNA) was identified exactly once per genome (see table 1).

Riboswitches

One-third of the known riboswitches are present in MAP JIII-386: Two copies of TPP and Cobalamin and one copy of SAM-IV, SAH and Glycine riboswitches, respectively, were found in all investigated *M. avium* genomes, see table 1. Additionally, two copies of SAH in MAH 104 were identified. The genome of CLIJ361 is lacking the TPP upstream of *thiE*.

Additionally, riboswitch features were found in several 5′-untranslated regions (UTRs); however, a function for these has not yet been confirmed: pan (synthesis of the vitamin pantothenate), pfl (absent in MAH 104), ydaO–yuaA, ykoK, and ykkC-III. The latter one has lost its second copies in MAP-C strains. Three of the riboswitches (SAM-IV, Cobalamin, and ykoK) have been reported previously in MAH, MAA (Ignatov et al. 2013), and *M. tuberculosis* (Arnvig and Young 2009) and were confirmed in this study.

The pan RNA motif represents a conserved RNA structure previously identified in only three bacterial families: *Chloroflexi*, *Firmicutes,* and *Proteobacteria* (Weinberg et al. 2010). Its secondary structure consists of one or two stem-loops containing two bulged adenosines and is located in 5′-UTRs of genes involved in the synthesis of the vitamin pantothenate. If the observed RNA motif is truly a pan-like sequence, it would be the first discovery of this RNA family in Actinobacteria.

Other ncRNAs

Using GORAP and manual alignment correction one PyrR-binding site was identified in each isolate, which is located upstream of a variety of genes involved in pyrimidine biosynthesis.

With the exception of MAP strain CLIJ361 (likely due to the limited assembly quality), one copy of 6C RNA was found within each investigated *M. avium* (Weinberg et al. 2007).

The Actino-pnp RNA motif was previously described as a conserved structure in Actinobacteria, apparently located in the 5′-UTR of genes encoding exoribonucleases (Weinberg et al. 2010). For each investigated *M. avium* strain, one copy of Actino-pnp was confirmed in this study (table 1).

The mraW RNA motif is a highly conserved RNA structure consisting of one hairpin with a highly conserved terminal loop sequence 5′-CUUCCCC-3′. Previously, it was predicted in many Actinobacteria and particularly within Mycobacteria. MraW was detected twice in investigated genomes, one copy being located consistently in the 5′-UTR of *mraW* genes and another copy, with similar secondary structure features, located in a region with multiple types of *mur* genes which likely form operons with *mraW.*

A study by Arnvig and Young (2009) discovered at least nine putative small RNA families in the genome of *M. tuberculosis* by RACE analysis and Northern blot experiments resulting in four *cis-* and five *trans*-encoded ncRNAs. With GORAP and a manual correction of Stockholm alignments, three of these ncRNAs were identified in all of the studied Mycobacteria samples: ASdes, ASpks, and F6; see table 1. Additionally, two ncRNA homologous classes were discovered: The *trans*-encoded ncRNA G2, which has been lost in MAP-C strains, and the AS1890 alignment, which achieved a very good bit score, however lacked the antisense protein homolog Rv1890c. These domains were described to act as *cis*-encoded and *trans*-encoded ncRNAs (Arnvig and Young 2009).

ASdes and ASpks are involved in lipid metabolism by regulating the polyketide synthase-12 (*pks12*) and fatty acid desaturase (*desA1*), respectively. The *pks12* gene contains two identical copies of ASpks, acting as antisense regulators of pks12 mRNA. In this study, two clusters of potential ASpks ncRNAs were identified: one cluster, including two identical copies of the region encoding ASpks, as described for *M. tuberculosis* (Arnvig and Young 2009) and a novel cluster, comprising one copy (in MAP-S) and two copies (in MAP-C and MAH 104). Within K-10′, *ASpks* homologs of the second cluster were detected in two copies, localized in different, but adjacent PKS genes: *pks7* and *pks8*. In addition to one copy of ASdes, located antisense of *desA1* gene, we were able to find further copies in *desA2*.

6S RNA is a highly abundant ncRNA, which was initially identified in *E. coli* (Hindley 1967) and was among the first small RNAs to be sequenced (Brownlee 1971), further believed to be necessary in at least one copy for each bacterium. By binding to the σ^{70}-containing housekeeping RNAP holoenzyme, it inhibits a large number of σ^{70}-dependent genes and thus enables a better adaption to stationary phase and environmental stress (Trotochaud and Wassarman 2004, 2006; Gildehaus et al. 2007; Cavanagh et al. 2008). Although 6S RNA is known for all bacteria branches (except Deinococcus/Thermus, Chlamydiae, most Actinobacteria) (Wehner, Damm, et al. 2014), until now no 6S RNA is known for Mycobacteria. Results of this study confirm these data: Based on the analysis of the eight investigated genome sequences, no 6S RNA could be identified.

Using GORAP, in all investigated MAP strains and MAH 104 about 80 ncRNAs were found (see table 1), whereas for *E. coli* about 155 and *S. entericus* about 200 ncRNAs could be detected (see supplementary table S20, Supplementary Material online). Some ncRNAs not known for the latter two bacteria were listed in supplementary table S19, Supplementary Material online. Possibly, in MAP strains and MAH strain 104 there are also more ncRNAs; however, they have not been studied intensively so far, and transcriptome profiles for discovering novel, specific ncRNAs are lacking and should be investigated in more detail in the future.

In 2013, Ignatov et al. described the noncoding transcriptome of *M. avium* resulting in 87 antisense and 10 intergenic small RNAs, which can roughly also be expected for MAP strains.

Altogether, in this study a different number for Aspks, G2, and YkkC-III among MAP Type-S and -C was detected.

Based on the multiple alignment of 70 ncRNAs, the phylogenetic reconstruction (fig. 4C) divides all MAP strains into MAP-S and MAP-C clusters, with a low bootstrap support within the highly similar MAP-C strains.

Loss/Gain of Gene Clusters

A lot of previous studies presented multiple LSPs between *M. avium* isolates including MAP-S and MAP-C (Type I/III and II) strains (Dohmann et al. 2003; Semret et al. 2005, Semret M, Turenne CY, de Haas P, et al. 2006; Marsh et al. 2006; Paustian et al. 2008; Alexander et al. 2009; Bannantine et al. 2012). Deletion or insertion of LSP regions was related to virulence properties of pathogenic bacteria and used as phylogenetic markers in *M. tuberculosis* complex (MTBC) strains (Alland et al. 2007). In *M. avium*, such regions encode also for metabolic enzymes and antigenic proteins; their specific distribution is an important source of genetic variability among members of the MAC (Semret M, Turenne CY, de Haas P, et al. 2006; Alexander et al. 2009). Within this study, not only the presence or absence of LSPs (consisting of at least four ORFs as gene cluster) but also the gain or loss of single genes was explored using comparative sequence analyses.

First, the presence or absence of 25 genome regions characteristically distributed between isolates of *M. avium* subspecies (Alexander et al. 2009) was confirmed in the examined genomes (see supplementary table S14*b*, Supplementary Material online). LSP^{A}11 was missing in MAP-S strain JIII-386 from Germany, as previously only reported for porcine MAP strain LN20 of sheep type originating from Canada (Alexander et al. 2009)—both belonging to MAP Type III.

Insertions

Ten regions of specific LSPs (LSP^{S}1–LSP^{S}10) present in MAP-S but absent in K-10 (Bannantine et al. 2012) were confirmed in S397 in this study and detected as homologous regions also in JIII-386 and CLIJ361 (table 3 and supplementary table S14*a* and *b*, Supplementary Material online). In contrast to Bannantine et al. (2012), four of these ten LSPs (LSP^{S}3, 6, 9, and 10) were identified also in all MAP-C strains. Furthermore, only LSP^{S}6 and LSP^{S}10 are absent in MAH 104. The distribution of the ten LSP^{S}s is MAP-type associated; it shows no differences between individual strains of MAP-S or MAP-C regarding their geographical origin.

However, our analysis showed, that LSP^{S}1 (9 kb) and LSP^{S}2 (6.6 kb) (Bannantine et al. 2012) are subsets of previously described larger elements LSP^{A}4-II (28.9 kb) and LSP^{A}18 (16.4 kb) identified in MAH 104 and MAP-S, but absent from MAP-C (Semret M, Turenne CY, de Haas P, et al. 2006). LSP^{A}4-II and LSP^{A}18 are related to the PIG-RDA20 and PIG-RDA10 regions detected by Dohmann et al. (2003). Based on the newly assembled JIII-386, homologous sequences of S397 and merged annotation, 23 ORFs (MAPs_15961-16180) homologous to LSP^{A}4-II sequences and comprising 8 ORFs of LSP^{S}1 were identified in the examined MAP-S strains. Two ORFs of LSP^{S}1 (MAPs_15940 and 15950) were absent in the genome of MAH 104 and no homologs were found in LSP^{A}4-II. This region could be extended by six adjacent ORFs (MAPs_15870–15930) and additionally by BacProt-annotated ORFs. A new LSP was defined: LSP^{S}Ia (see table 4) comprising 11 ORFs, absent in MAP-C and absent in MAH 104. This LSP really could represent an insertion, with genes encoding proteins involved in CoA energy metabolism and tetracycline-controlled transcriptional activation. Furthermore, LSP^{S}2 (MAPs_46190–46270) matched to LSP^{A}18, the nine ORFs of LSP^{S}2 are homologous to ORFs

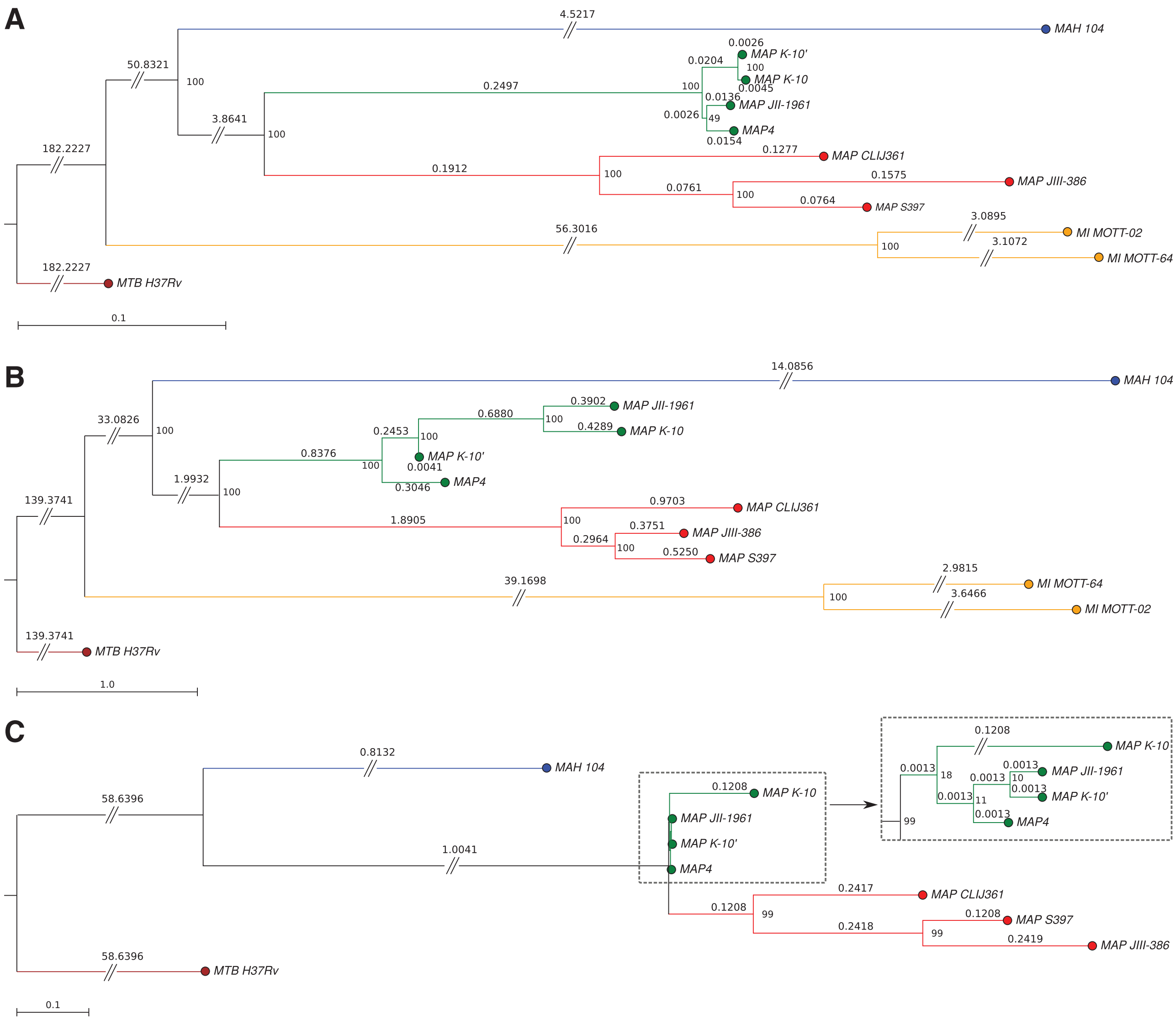

Fig. 4.—Phylogenetic reconstructions for all investigated *M. avium* strains based on sequence comparison of 790 corresponding CDSs on nucleotide (*A*) and amino acid level (*B*) and 70 corresponding ncRNAs (*C*). *Mycobacterium tuberculosis* strain H37Rv was used as an outgroup. *Mycobacterium intracellulare* (MI) strains were included as members of the MAC. Float numbers correspond to substitutions per site and integer numbers represent RAxML bootstrap values. Long branches are shrinked. Detailed figures, all multiple sequence alignments, and tree representations in Newick format can be found in the supplementary figures S22–S26, Supplementary Material online. Strains of MAP-S, Type I: CLIJ361 and Type III: JIII-386, S397 (red); Strains of MAP-C, Type II: K-10, K-10′, MAP4, JII-1961 (green); MAH strain 104 (blue); MI MOTT-64 and MI MOTT-02 (orange); and *M. tuberculosis* strain H37Rv (brown, used as outgroup) are shown.

MAV5227–5235 in MAH 104. LSP^{S}2 was combined with LSP^{S}4, extended by ten adjacent ORFs and newly designated as LSP^{S}II (see table 4).

ORFs belonging to LSP^{S}5 and LSP^{S}7 (see table 3), and additionally six adjacent ORFs (MAPs_17621, MAPs_17622, MAPs_17680–17710) were described as novel region in MAP-S and MAH 104 genomes by Bannantine et al. (2012) comprising also the GPL region (missing MAPs_17680-17710) published by Alexander et al. (2009). This genome region was predicted to encode proteins involved in the biosynthesis of glycopeptidolipids (GPLs) (Eckstein et al. 2003). GPLs are discussed to contribute to the virulence of members of the MAC. Different genes involved in the synthesis of GPLs would be expected to alter indirectly the interaction of the bacterium with its host. We analyzed that MAPs_17650, 17670, and 17690 are homologous to the GPL genes *mtfC*, *dhgA*, and *hlpA* belonging to the GPL biosynthesis cluster that is known to be diversely organized among individual strains and subspecies of *M. avium* (Eckstein et al. 2003; Krzywinska and Schorey 2003). In this study the 14 ORFs were detected to

Table 3
Distribution of Ten LSPs in *Mycobacterium avium* Strains, Previously Described to be Present in MAP-S but Absent in MAP-C

Name	Size (kb)	ORFs	MAP-C				MAP-S			MAH
			K-10	K-10′	MAP4	JII-1961	JIII-386	S397	CLIJ361	104
LSP^{S}1	9.01	MAPs_15940–16060	—	—	—	—	Full	Full	Full[a]	Part
LSP^{S}2	6.65	MAPs_46190–46270	—	—	—	—	Full	Full	Full[a]	Full[a]
LSP^{S}3	3.78	MAPs_14620–14660	Full	Full	Full	Full	Full	Full	Full[a]	Full
LSP^{S}4	3.63	MAPs_46290–46320	—	—	—	—	Full	Full	Full[a]	Full
LSP^{S}5	3.47	MAPs_17580–17610	—	—	—	—	Full	Full	Full	Part
LSP^{S}6	3.0	MAPs_40470–40500	Full	Full	Full	Full	Full	Full	Full	—
LSP^{S}7	2.89	MAPs_17640–17670	—	—	—	—	Full	Full	Full	Full
LSP^{S}8	2.39	MAPs_02730–02760	Part	Part	Part	Part	Full	Full	Full[a]	Full
LSP^{S}9	1.84	MAPs_23120–23150	Full	Full	Full	Full	Full	Full	Full	Full
LSP^{S}10	1.58	MAPs_42460–42490	Full	Full	Full	Full	Full	Full	Full	—

Note.—Labels and locations according to Bannantine et al. (2012). LSP^{S}8 was only partially detected with an alignment length of 692 bp in all MAP-C strains. Homologous fasta sequences for LSP^{S}1–10 of MAP JIII-386 as well as further details and additional information about the distribution of 25 other LSPs (Semret et al. 2005; Alexander et al. 2009) can be found in the supplementary table S14*a* and *b*, Supplementary Material online. Full, full-length hit; Part, partial hit.

[a] All ORFs comprised by the LSP^{S} are present but split on different contigs or genomic locations.

be present in MAP-S, and absent in the examined MAP-C strains. It was possible to assign a function for MAPs_17620, 17621, and 17690 (see supplementary table S17, Supplementary Material online). Otherwise, MAPs_17690–17710 are absent in MAH 104, but genes *dhgA* and *mtfC* are still present in the annotation of MAH 104. Altogether, this region included in addition five BacProt-annotated ORFs and was newly designated as LSP^{S}III (see table 4).

A further region (21.3 kb) was identified and newly defined as LSP^{S}IV, comprising 22 ORFs (MAPs_20550–20770). This LSP is present in JIII-386, S397, CLIJ361 and MAH 104 (with 243 mismatches), but absent in MAP-C strains (see table 4). Additionally, in JIII-386 two ORFs were predicted on the opposite strand by BacProt. Sequences of 15 ORFs (MAPs_20620–20770) are homologous to sequences of previously described LSP MAV-14 (Alexander et al. 2009); seven adjacent ORFs of LSP^{S}IV (MAPs_20550-20610) are absent from MAV-14.

Deletions

Several deletions in MAP-S strains, which have already been described earlier, were verified in this study, but also differences were found. Three gene clusters (LSPs) comprising 32 genes, annotated in MAP K-10 were previously characterized to be absent in MAP-S isolates: MAP1432–MAP1438c (deletion sΔ-1), MAP1484c–MAP1491 (deletion #1), and MAP1728c–MAP1744 (deletion #2) (Marsh and Whittington 2005; Marsh et al. 2006; Semret M, Turenne CY, de Haas P, et al. 2006; Paustian et al. 2008; Bannantine et al. 2012). Genes included in deletion sΔ-1, deletion #1, and deletion #2 were tested to be absent in sheep strains from the United States (Paustian et al. 2008; Bannantine et al. 2012); those of deletion #1 and #2 tested as absent in Australian sheep strains (Marsh et al. 2006). In this study, genes of deletions #1 and #2 were also absent in JIII-386 from Germany and CLIJ361 from Australia. But in contrast to sheep strain S397 from the United States, the seven K-10 genes belonging to deletion sΔ-1 were identified as being present in sheep strains JIII-386 from Germany and CLIJ361 from Australia (genes MAP1433c–MAP1438c in full length; MAP1432 with mismatches; see table 5 and supplementary table S16, Supplementary Material online). Differences regarding the presence or absence of deletion sΔ-1 could reflect diversities among MAP-S strains originating from different geographic regions of the world. Marsh et al. (2006) identified ORF MAP2325 in cattle strains but its loss (designated as deletion #3) in Australian sheep isolate Telford 9.2 (MAP-S, Type I) using microarray and confirmed this deletion #3 in 16 sheep strains by PCR. In contrast, MAP2325 was found to be present in MAP-S (Type III) isolates from the United States (Paustian et al. 2008; Bannantine et al. 2012). This discrepancy suggested a difference between MAP isolates recovered from sheep in Australia and the United States. Furthermore, within this study, MAP2325 could be found with 100% sequence identity also in MAP-S strains JIII-386 (Type III) from Germany as well as in CLIJ361 (Type I) from Australia, and confirmed in all MAP-C isolates. Again, results reflect diversities within MAP-S group, but could also partially indicate discrepancies between results of different methods (sequencing, microarray, and PCR).

Loss and Gain of Genes

Based on the merged annotations between NCBI and BacProt, new differences regarding the loss and gain of single protein-coding genes (CDSs) among MAP-S and MAP-C strains were detected. In JIII-386, S397 and CLIJ361 genomes, 80 homologous CDSs were identified which are absent from K-10/K-10′ and 82 homologous CDSs which are absent from MAP4 and

Table 4

New LSPs Regions, Extended and Revised Previous Described Regions, Present in MAP-S but Absent in MAP-C

LSP (Genomic Region)	LSP^{S} Included[a]	New LSP	Island Size (bp) (MAP-C negative)	Including MAPs	# ORFs	# ORFs (BacProt)	Present in		
							MAP-S	MAP-C	MAH 104
		LSP^{S}I (a+b)	34,377	**MAPs_15870–16180**	31	+5	Yes	Not	Partly
New (This Study)	2 ORFs of LSP^{S}1[a]	LSP^{S}Ia	10,227	MAPS_15870–15950	9	+2	Yes	Not	Not
Extended (This Study)	23 ORFs of LSP^{A}4-II[b]	LSP^{S}Ib	24,150	MAPS_15961–16180	22	+3	Yes	Not	Yes
	8 ORFs of LSP^{S}1								
Extended (This Study)	LSP^{S}2 + 4[a]/LSP^{A}18[b]	**LSP^{S}II**	16,392	**MAPs_46170–46350**	18	+2	Yes	Not	Yes
	New (BacProt)			MAPs_46241–46242[c]	1				
Extended (Previously[a])	LSP^{S}5 + 7[a]/GPL[b]	**LSP^{S}III (a+b)**	16,015	**MAPs_17580–17700**	12	+5	Yes	Not	Partly
	New (BacProt)			MAPs_17690[c]	1		Yes	Not	Not
	LSP^{S}5 + 7[a]/GPL[b]	LSP^{S}IIIa	12,142	MAPS_17580–17680	11	+4	Yes	Not	Yes
		LSP^{S}IIIb	3,873	MAPS_17690–17700	2	+1	Yes	Not	Not
Extended (This Study)	MAV-14[b]	**LSP^{S}IV**	≥21,310	**MAPs_20550–20770**	22	+2	Yes	Not	Yes
Revised (This Study)	LSP^{S}3			—[a]	—[a]		Yes	Yes	Yes
	LSP^{S}6			—[a]	—[a]		Yes	Yes	Not
	LSP^{S}8			—[a]	—[a]		Yes	Partly	Yes
	LSP^{S}9			—[a]	—[a]		Yes	Yes	Yes
	LSP^{S}10			—[a]	—[a]		Yes	Yes	Not

NOTE.—The number of novel ORFs, additionally predicted by BacProt and with no overlap against previously annotated MAPs ORFs, is listed. For further information about genomic positions of homologous ORFs (CDSs) in MAP JIII-386 and gene annotation, see supplementary table S17, Supplementary Material online. # ORFs, number of ORFs including homologous as well as hypothetical ORFs. In bold: new designation of LSPs and the included MAPs.

[a] See Bannantine et al. (2012).

[b] See Alexander et al. (2009).

[c] BacProt-assigned function.

JII-1961. Supplementary table S15*a*, Supplementary Material online, presents in detail gain and loss of genes detected in this study comparing MAP genomes and MAH 104. The 40 genes with assigned functions (homologous genes) as well as the 30 hypothetical genes, whose were previously described by Bannantine et al. (2012) to be present in three sheep isolates of MAP-S, Type III (from the United States) but absent in K-10 strain (MAP-C), were part of this analysis. However, 36 of these 70 genes belonged to the 10 MAP-S-specific LSP^{S} regions also published by Bannantine et al. (2012) including all ORFs of LSP^{S}1, 2, 4, 5 and 7, and two of four ORFs of LSP^{S}8 (see table 3). Four genes (hypothetical genes) were still present in MAP-C strains (supplementary table S15*b*, Supplementary Material online). For 9 of the 30 above-mentioned hypothetical genes, it was possible to assign a function based on homology. In total, 34 additional ORFs were found in all MAP-S, but absent in MAP-C, among them 5 ORFs which were annotated only by BacProt (supplementary table S17, Supplementary Material online).

Altogether 80 CDSs (ORFs with an assigned function), present in MAP-S but absent in MAP-C strains, were annotated in this study and listed in supplementary table S17, Supplementary Material online. Nine CDSs were also absent in MAH 104. Eight of these genes belong to the new designated LSP^{S}Ia, possibly indicating a specific insertion region into MAP-S strains.

MAP-S (Type III) strains JIII-386 and S397 differed in the presence and/or absence of altogether 33 CDS (see supplementary table S15*a*, Supplementary Material online). In detail, 25 CDSs of S397 were present in MAP-C and partially in CLIJ361 but absent in JIII-386 including four ORFs (MAPs_23210–23240), and six ORFs (MAPs_39450–39500), possibly representing specific deletions in JIII-386. The last gene cluster encodes for three mammalian cell entry (mce) family proteins and virulence factor mce. *Mce* genes were originally identified and studied in *M. tuberculosis* and have been associated with survival within macrophages and increased virulence in this species (see review of Paustian et al. 2010). Eight CDSs of JIII-386 were present in MAP-C, CLIJ361, and MAH 104 but absent in S397 and included four complete ORFs with an assigned function (CDSs) of deletion sΔ-1. In contrast, MAP-C type strains showed high similarities regarding their gene repertoire. Only two genes are absent from MAP4 (coding for ATP/GTP-binding integral membrane protein and CsbD-like protein) and two other genes are absent from JII-1961 (coding for inosine 5-monophosphate dehydrogenase and a PE-PGRS family protein) (see supplementary table S15*a*, Supplementary Material online). As expected, with loss and gain of about 700 CDSs, MAH strain 104 emerged as the most different strain among the investigated Mycobacteria (see supplementary table S15*a*, Supplementary Material online). The large number of genes (up to 208 compared with K-10′), absent in MAP-S, Type I strain CLIJ361, includes a high amount of false-negative hits most likely caused by the lower assembly quality. Probably, some of the genes are present in the genome of CLIJ361 but could not be

Table 5

Gene Cluster Comprising Seven K-10 ORFs Absent in S397 but Present in JIII-386 and CLIJ361

ORF	Size (bp)	Description
MAP1432[a]	1,490	REP-family protein
MAP1433c[b]	1,745	3-oxosteroid 1-dehydrogenase
MAP1434[b]	1,118	Putative phthalate oxygenase
MAP1435	713	Short chain dehydrogenase
MAP1436c[b]	782	Putative oxidoreductase
MAP1437c	986	Hypothetical protein[c]
MAP1438c[d]	983	Probable lipase[c]

NOTE.—Table based on Bannantine et al. (2012). Homologous sequences of all ORFs were found on scaffold S02 in MAP JIII-386. For additional information and BLAST results, see supplementary table S16, Supplementary Material online.

[a]Partial hit (alignment length 1,484 bp) with mismatches.

[b]Involved in energy metabolism.

[c]Treated as hypothetical ORFs during analyses.

[d]Involved in degradation of macromolecules.

identified in nearly full-length and were therefore counted as absent from this strain. Nevertheless, better assembled MAP Type I strains could enable more reliable comparisons among MAP-S: Type I and III strains.

PE/PPE/PGRS Genes and mmpL5

The PE and PPE gene families are restricted to mycobacteria, encode acidic, glycine-rich proteins and several of them are proposed to be involved in antigenic variation and in the pathogenesis of infection (Ramakrishnan et al. 2000; Dubnau et al. 2002; Li et al. 2005). They comprise anywhere from 1% of the genome (MAP) to nearly 10% (*M. tuberculosis*) (Paustian et al. 2010). In this study individual strains show 6, 7, or 8 PE genes as well as 32 (JIII-386, S397) or 33 and 35 (MAP-C strains) PPE genes, annotated by BacProt—there is no clear differentiation between MAP-S and MAP-C strains possible. Furthermore, PE-PGRS family protein genes—the largest subfamily of PE family genes, also suggested to play an important role in the persistence of mycobacteria and to be involved in antigenic variation and immune evasion (Tian and Jian-Ping 2010)—were searched. It was previously assumed that *M. avium*, including MAH and MAP lacks these PE-PGRS family protein genes (Brennan and Delogu 2002; Cole 2002; Li Y, Miltner E, et al. 2005; Delogu et al. 2006; van Pittius et al. 2006). However, in this study at least one (S397, CLIJ361, MAH 104) or two (JIII-386, K-10′, JII-1961, MAP4) homolog to PE-PGRS gene family could be annotated by BacProt, confirming results of Tian and Jian-Ping (2010) for *M. avium*. Marri et al. (2006) suggested that the paucity of PE/PPE virulence genes in MAP in comparison to *M. tuberculosis* was compensated by the acquisition of other virulence factors as a result of lateral gene transfer.

Otherwise, many mycobacterial membrane protein large (mmpL) genes are associated with clusters involved in the biosynthesis of cell wall-associated glycolipids (Cole et al. 1998). *MmpL5* gene encodes a protein involved in lipid transport (Marri et al. 2006). This study confirms the absence of *mmpL5* gene in MAP-S strains and its presence in MAP-C strains (and MAH 104) previously described by Marsh and Whittington (2005) possibly indicates that some of these *mmpL* gene products could also help in host association.

Single Nucleotide Variants

Depending on MAP type and *M. avium* subspecies, different numbers of SNVs were detected within corresponding CDS, see supplementary table S18, Supplementary Material online. Among MAP-C strains less than 200 SNVs and among MAP-S strains about 1,000 SNVs were identified. As obvious from the genome comparison (for JIII-386 and S397 shown in fig. 2, not shown for MAP-C strains), these results confirm a higher heterogeneity within the MAP-S group and high similarity between MAP-C strains. This could indicate that MAP-C has evolved more recently or over a long time within a restricted niche. Among CDSs of MAP-C and MAP-S, Type III strains more than 2,000 SNVs were found. More than 26,000 SNVs were detected comparing CDSs of MAP Type I, II and III strains with MAH 104, revealing the high evolutionary distance between MAP and MAH. As shown before for the MTBC (Hershberg et al. 2008) also two-thirds of SNVs among MAP strains are nonsynonymous (see supplementary table S18, Supplementary Material online) which is unlike in most other organisms in which synonymous SNVs predominate. This has been proposed to be the consequence of the relatively short evolutionary age of MTBC (Stucki and Gagneux 2013) which applies also to MAP. Furthermore, this could indicate an adaptive evolution of MAP to different hosts with positive selective pressure (Hsu et al. 2011).

Phylogenetic Reconstruction/Ancestral Relationship

Together with the other members of *M. avium* (MAH, MAA, and MAS) and the genetically related *Mycobacterium intracellulare*, MAP belongs to the MAC revealing different pathogenicity and infecting different hosts. How MAP has evolved into a professional pathogen of ruminants remains largely unknown, also its division into two main lineages: MAP-S (MAP Type I/III) and MAP-C (Type II). Previously, two models for a putative biphasic evolution of MAP to MAP-S and MAP-C strains were proposed. In model I, the first phase is characterized by the emergence of an original pathogenic clone of MAP (proto-MAP) from a strain of MAH through acquisition of novel DNA and polymorphisms shared by all modern strains (Alexander et al. 2009). The second phase includes the subsequent differentiation from proto-MAP to sheep and cattle lineages (MAP-S and MAP-C) through genomic insertion/or deletion of different LSPs (Alexander et al. 2009). Model II suggests that a different number of independent inversion events and loss of LSPs causes the evolution from MAH or from *M. intracellulare* to proto-MAP and further through sheep type to cattle type (Bannantine et al. 2012).

In this study, the detected higher number of individual CDSs in MAP-S as in MAP-C and the higher number of sequence regions present in MAP-S and MAH 104 but absent in MAP-C (see LSP^{S}s) than deletions in MAP-S support the model of evolution from proto-MAP through sheep type to cattle type. Furthermore, the higher diversity among MAP-S strains (see SNVs) could also indicate an evolutionary earlier onset of MAP-S in comparison to MAP-C. Otherwise, the calculated phylogenetic trees (fig. 4 and supplementary figs. S22–S26, Supplementary Material online)—based on comparison of nucleotide- or amino acid sequences within 790 corresponding CDSs and additionally of corresponding ncRNA sequences—give ambiguous results regarding proposed evolutionary models (fig. 4). The trees illustrate the large genetic distance between MAP, MAH, and *M. intracellulare* and they show clearly that the subspecies MAP is more closely related to MAH than to *M. intracellulare* (fig. 4*A* and *B*). Depending on the type of compared sequences, trees exhibit different results concerning higher or lower similarity of MAP-S or MAP-C to MAH (fig. 4*A* vs. fig. 4*B* and *C*). Consequently, results of current phylogenetic trees do not answer the question whether MAP-S is the evolutionary intermediate between proto-MAP and MAP-C or whether there was another way of division into the two main lineages during MAP evolution.

Summarizing several previous studies also provides ambiguous results contradicting both proposed evolutionary models. Sohal et al. (2010) described that SNPs in IS*1311* could be indicative of the MAP-S type being an evolutionary intermediate between *M. avium* and MAP-C type, but SNPs in the *hsp65* gene (Turenne et al. 2006) indicate that MAP-C is the intermediate. Otherwise, Marsh and Whittington (2007) identified 11 SNPs between MAP-S and -C strains in 8 genes, all present in MAH 104 and distributed almost evenly among both MAP types. Furthermore, there are polymorphic regions unique to MAP-S strains and MAH 104, but also large deletions in the MAP-S strains (Marsh et al. 2006).

To decipher the complexity of the evolutionary processes leading to MAP-S and MAP-C strains, future genome comparisons should investigate additional target regions or genes as such for metabolic pathways, and especially use a higher number of WGS of MAP-S, Type I and III as well as of related strains among the species *M. avium* (MAA, MAS, and MAH).

Conclusion

With the newly sequenced JIII-386 genome the so far best assembled MAP-S sequence was presented here, although still representing a draft genome. Using merged results from NCBI and BacProt a comprehensive annotation of CDSs was obtained, including a large fraction of CDSs identified by both approaches, and also additional ones identified exclusively with one of the approaches. This relativizes absolute numbers of annotated genes in studies using only one annotation program. Newly annotated CDSs complete the previously detected differences between MAP-S and MAP-C strains. Within this study, BacProt reannotations of CDSs for each of the seven *M. avium* strains are provided. A new Shine–Dalgarno sequence motif was extracted; further studies should disclose whether this motif was conserved among Mycobacteria.

For the first time about 80 ncRNAs and riboswitches of MAP were presented, differing in numbers in three cases from MAH 104 but also between MAP-S and -C. Furthermore, a pan-like sequence was observed, which is the first discovery of this RNA family in Actinobacteria. The performed genome comparison is the most comprehensive comparison to date since it comprises three MAP-S and three MAP-C isolates from three, respectively, two different continents. Using extended annotation, previously reported genome differences between S and C strains were partially revised and new MAP Type-S-specific regions were identified.

The concordant presence and absence of specific LSPs and distribution of ncRNAs among the examined MAP-S, Type I and III strains show that these strains are very closely related subgroups of MAP-S.

In conclusion, our data will improve the understanding of the MAP genome, help to decipher the genetic basis for different phenotypic characteristics of MAP-S and -C (Type I/III and II, respectively) strains and the evolution of MAP types, also in relation to MAH.

Supplementary Material

Supplementary tables S1–S21 and figures S10*a* and *b* and S22–S26 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Kerstin Steger (IMP, Friedrich-Loeffler-Institut, Jena), Stefanie Thies (Helmholtz Centre for Infection Research, Braunschweig), and Ivonne Heinze and Ivonne Görlich (Leibniz Institute for Age Research—Fritz-Lipman-Institute, Jena) for their excellent technical assistance. This work was supported in part by DFG MA5082/1-1 (to M.M. and M.H.), DFG SPP-1569 (to M.H.), and the Carl-Zeiss-Stiftung (M.M.).

Literature Cited

Alexander DC, Turenne CY, Behr MA. 2009. Insertion and deletion events that define the pathogen *Mycobacterium avium* subsp. *paratuberculosis*. J Bacteriol. 191(3):1018–1025.

Alland D, et al. 2007. Role of large sequence polymorphisms (LSPs) in generating genomic diversity among clinical isolates of *Mycobacterium tuberculosis* and the utility of LSPs in phylogenetic analysis. J Clin Microbiol. 45(1):39–46.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403–410.

Amonsin A, et al. 2004. Multilocus short sequence repeat sequencing approach for differentiating among *Mycobacterium avium* subsp. *paratuberculosis* strains. J Clin Microbiol. 42(4):1694–1702.

Arnvig KB, Young DB. 2009. Identification of small RNAs in *Mycobacterium tuberculosis*. Mol Microbiol. 73(3):397–408.

Arnvig KB, Young DB. 2010. Regulation of pathogen metabolism by small RNA. Drug Discov Today Dis Mech. 7(1):e19–e24.

Atreya R, et al. 2014. Facts, myths and hypotheses on the zoonotic nature of *Mycobacterium avium* subspecies *paratuberculosis*. Int J Med Microbiol. 304(7):858–867.

Bankevich A, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19(5):455–477.

Bannantine JP, et al. 2012. Genome sequencing of ovine isolates of *Mycobacterium avium* subspecies *paratuberculosis* offers insights into host association. BMC Genomics 13(1):89.

Bannantine JP, et al. 2014. Complete genome sequence of *Mycobacterium avium* subsp. *paratuberculosis*, isolated from human breast milk. Genome Announc. 2(1):e01252–13.

Barrick JE, et al. 2004. New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. Proc Natl Acad Sci U S A. 101(17):6421–6426.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27(4):578–579.

Brennan MJ, Delogu G. 2002. The PE multigene family: a 'molecular mantra' for mycobacteria. Trends Microbiol. 10(5):246–249.

Brownlee GG. 1971. Sequence of 6s RNA of *E.* coli. Nat New Biol. 229(5):147–149.

Castellanos E, et al. 2009. Discovery of stable and variable differences in the *Mycobacterium avium* subsp. *paratuberculosis* type i, ii, and iii genomes by pan-genome microarray analysis. Appl Environ Microbiol. 75(3):676–686.

Castellanos E, et al. 2010. Molecular characterization of *Mycobacterium avium* subspecies *paratuberculosis* types II and III isolates by a combination of MIRU–VNTR loci. Vet Microbiol. 144(1):118–126.

Cavanagh AT, Klocko AD, Liu X, Wassarman KM. 2008. Promoter specificity for 6S RNA regulation of transcription is determined by core promoter sequences and competition for region 4.2 of sigma70. Mol Microbiol. 67(6):1242–1256.

Chu TC, et al. 2013. Assembler for *de novo* assembly of large genomes. Proc Natl Acad Sci U S A. 110(36):E3417–E3424.

Clarke C, Little D. 1996. The pathology of ovine paratuberculosis: gross and histological changes in the intestine and other tissues. J Comp Pathol. 114(4):419–437.

Cole S, et al. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393(6685):537–544.

Cole ST. 2002. Comparative and functional genomics of the *Mycobacterium tuberculosis* complex. Microbiology 148(10):2919–2928.

Collins DM, Gabric DM, de Lisle GW. 1990. Identification of two groups of *Mycobacterium paratuberculosis* strains by restriction endonuclease analysis and DNA hybridization. J Clin Microbiol. 28(7):1591–1596.

Dale J, et al. 1995. Mobile genetic elements in *Mycobacteria*. Eur Respir J Suppl. 20:633s–648s.

Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14(7):1394–1403.

Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5(6):e11147.

De Juan L, et al. 2006. Molecular epidemiology of types I/III strains of *Mycobacterium avium* subspecies *paratuberculosis* isolated from goats and cattle. Vet Microbiol. 115(1):102–110.

De Juan L, Mateos A, Dominguez L, Sharp J, Stevenson K. 2005. Genetic diversity of *Mycobacterium avium* subspecies *paratuberculosis* isolates from goats detected by pulsed-field gel electrophoresis. Vet Microbiol. 106(3):249–257.

Delogu G, et al. 2006. PE_PGRS proteins are differentially expressed by *Mycobacterium tuberculosis* in host tissues. Microbes Infect. 8(8):2061–2067.

Dohmann K, et al. 2003. Characterization of genetic differences between *Mycobacterium avium* subsp. *paratuberculosis* type i and type ii isolates J Clin Microbiol. 41(11):5215–5223.

Dubnau E, Fontán P, Manganelli R, Soares-Appel S, Smith I. 2002. *Mycobacterium tuberculosis* genes induced during infection of human macrophages. Infect Immun. 70(6):2787–2795.

Eckstein TM, Belisle JT, Inamine JM. 2003. Proposed pathway for the biosynthesis of serovar-specific glycopeptidolipids in *Mycobacterium avium* serovar 2. Microbiology 149(10):2797–2807.

Englund S, Bölske G, Ballagi-Pordany A, Johansson KE. 2001. Detection of *Mycobacterium avium* subsp. *paratuberculosis* in tissue samples by single, fluorescent and nested PCR based on the IS900 gene. Vet Microbiol. 81(3):257–271.

Fritsch I, Luyven G, Köhler H, Lutz W, Möbius P. 2012. Suspicion of *Mycobacterium avium* subsp. *paratuberculosis* transmission between cattle and wild-living red deer (*Cervus elaphus*) by multitarget genotyping. Appl Environ Microbiol. 78(4):1132–1139.

Gardner PP, et al. 2009. Rfam: updates to the RNA families database. Nucleic Acids Res. 37(Suppl 1):D136–D140.

Ghosh P, et al. 2012. Genome-wide analysis of the emerging infection with *Mycobacterium avium* subspecies *paratuberculosis* in the Arabian camels (*Camelus dromedarius*). PLoS One 7(2):e31947.

Gildehaus N, Neusser T, Wurm R, Wagner R. 2007. Studies on the function of the riboregulator 6S RNA from *E. coli*: RNA polymerase binding, inhibition of in vitro transcription and synthesis of RNA-directed de novo transcripts. Nucleic Acids Res. 35(6):1885–1896.

Green E, et al. 1989. Sequence and characteristics or IS900, an insertion element identified in a human Crohn's disease isolate of *Mycobacterium paratuberculosis*. Nucleic Acids Res. 17(22):9063–9073.

Griffiths-Jones S. 2005. RALEERNA ALignment editor in Emacs. Bioinformatics 21(2):257–259.

Hershberg R, et al. 2008. High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. PLoS Biol. 6(12):e311.

Hindley J. 1967. Fractionation of 32p-labelled ribonucleic acids on polyacrylamide gels and their characterization by fingerprinting. J Mol Biol. 30(1):125–136.

Hsu CY, Wu CW, Talaat AM. 2011. Genome-wide sequence variation among *Mycobacterium avium* subspecies *paratuberculosis* isolates: a better understanding of Johne's disease transmission dynamics. Front Microbiol. 2(236):1–14.

Ignatov D, et al. 2013. RNA-Seq analysis of *Mycobacterium avium* noncoding transcriptome. PLoS One 8(9):e74209.

Junier T, Zdobnov EM. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics 26(13):1669–1670.

Katoh K, Misawa K, Kuma KI, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30(14):3059–3066.

Krzywinska E, Schorey JS. 2003. Characterization of genetic differences between *Mycobacterium avium* subsp. *avium* strains of diverse virulence with a focus on the glycopeptidolipid biosynthesis cluster. Vet Microbiol. 91(2):249–264.

Lagesen K, et al. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35(9):3100–3108.

Lechner M. 2009. Detection of orthologs in large-scale analysis [Master's thesis]. Germany: University of Leipzig.

Lechner M, et al. 2011. Proteinortho: detection of (Co-)orthologs in large-scale analysis. BMC Bioinformatics 12(1):124.

Lechner M, et al. 2014. Genomewide comparison and novel ncRNAs in Aquificales. BMC Genomics 15(1):522.

Li L, Bannantine JP, et al. 2005. The complete genome sequence of *Mycobacterium avium* subspecies *paratuberculosis*. Proc Natl Acad Sci U S A. 102(35):12344–12349.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659.

Li Y, Miltner E, Wu M, Petrofsky M, Bermudez LE. 2005. A *Mycobacterium avium* PPE gene is associated with the ability of the bacterium to grow in macrophages and virulence in mice. Cell Microbiol. 7(4):539–548.

Li Z, et al. 2012. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de–bruijn–graph. Brief Funct Genomics. 11(1):25–37.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25(5):955–964.

Mandal M, et al. 2004. A glycine-dependent riboswitch that uses cooperative binding to control gene expression. Science 306(5694):275–279.

Marri PR, Bannantine JP, Paustian ML, Golding GB. 2006. Lateral gene transfer in *Mycobacterium avium* subspecies *paratuberculosis*. Can J Microbiol. 52(6):560–569.

Marsh I, Whittington R. 2005. Deletion of an *mmpL* gene and multiple associated genes from the genome of the s strain of *Mycobacterium avium* subsp. *paratuberculosis* identified by representational difference analysis and in silico analysis. Mol Cell Probes. 19(6):371–384.

Marsh I, Whittington R. 2007. Genomic diversity in *Mycobacterium avium*: single nucleotide polymorphisms between the S and C strains of *M. avium* subsp. *paratuberculosis* and with M. a. avium. Mol Cell Probes. 21(1):66–75.

Marsh IB, et al. 2006. Genomic comparison of *Mycobacterium avium* subsp. *paratuberculosis* sheep and cattle strains by microarray hybridization. J Bacteriol. 188(6):2290–2293.

Mijs W, et al. 2002. Molecular evidence to support a proposal to reserve the designation *Mycobacterium avium* subsp. *avium* for bird-type isolates and '*M. avium* subsp. *hominissuis*' for the human/porcine type of M. *avium*. Int J Syst Evol Microbiol. 52(5):1505–1518.

Miotto P, et al. 2012. Genome-wide discovery of small RNAs in *Mycobacterium tuberculosis*. PLoS One 7(12):e51950.

Möbius P, Fritsch I, Luyven G, Hotzel H, Köhler H. 2009. Unique genotypes of *Mycobacterium avium* subsp. *paratuberculosis* strains of Type III. Vet Microbiol. 139(3):398–404.

Möbius P, Luyven G, Hotzel H, Köhler H. 2008. High genetic diversity among *Mycobacterium avium* subsp. *paratuberculosis* strains from German cattle herds shown by combination of IS900 restriction fragment length polymorphism analysis and mycobacterial interspersed repetitive unit-variable-number tandem-repeat typing. J Clin Microbiol. 46(3):972–981.

Nahvi A, et al. 2002. Genetic control by a metabolite binding mRNA. Chem Biol. 9(9):1043–1049.

Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: inference of RNA alignments. Bioinformatics 25(10):1335–1337.

Olsen I, Johansen TB, Billman-Jacobe H, Nilsen SF, Djønne B. 2004. A novel IS element, *ISMpa1*, in *Mycobacterium avium* subsp. *paratuberculosis*. Vet Microbiol. 98(3):297–306.

Over K, Crandall PG, O'Bryan CA, Ricke SC. 2011. Current perspectives on *Mycobacterium avium* subsp. *paratuberculosis*, Johne's disease, and Crohn's disease: a review. Crit Rev Microbiol. 37(2):141–156.

Papenfort K, Vogel J. 2010. Regulatory RNA in bacterial pathogens. Cell Host Microbe 8(1):116–127.

Paustian M, et al. 2008. Comparative genomic analysis of *Mycobacterium avium* subspecies obtained from multiple host species. BMC Genomics 9(1):135.

Paustian ML, Bannantine JP, Kapur V. 2010. *Mycobacterium avium* subsp. *paratuberculosis* genome. Chapter 8. In: Behr MA, Collins DM, editors. Paratuberculosis: organism, disease, control. Oxford: CAB International p. 73–81.

Pickup R, et al. 2006. *Mycobacterium avium* subsp. *paratuberculosis* in lake catchments, in river water abstracted for domestic use, and in effluent from domestic sewage treatment works: diverse opportunities for environmental cycling and human exposure. Appl Environ Microbiol. 72(6):4067–4077.

Ramakrishnan L, Federspiel NA, Falkow S. 2000. Granuloma-specific expression of *Mycobacterium* virulence proteins from the glycine-rich PE-PGRS family. Science 288(5470):1436–1439.

Rhodes G, et al. 2014. *Mycobacterium avium* subspecies *paratuberculosis*: human exposure through environmental and domestic aerosols. Pathogens 3(3):577–595.

Rindi L, Garzelli C. 2014. Genetic diversity and phylogeny of *Mycobacterium avium*. Infect Genet Evol. 21:375–383.

Sachse K, et al. 2014. Evidence for the existence of two new members of the family *Chlamydiaceae* and proposal of *Chlamydia avium* sp. nov. and *Chlamydia gallinacea* sp. nov. Syst Appl Microbiol. 37(2):79–88.

Semret M, et al. 2005. Genomic polymorphisms for *Mycobacterium avium* subsp. *paratuberculosis* diagnostics. J Clin Microbiol. 43(8):3704–3712.

Semret M, Turenne CY, Behr MA. 2006. Insertion sequence IS900 revisited. J Clin Microbiol. 44(3):1081–1083.

Semret M, Turenne CY, de Haas P, Collins DM, Behr MA. 2006. Differentiating host-associated variants of *Mycobacterium avium* by PCR for detection of large sequence polymorphisms. J Clin Microbiol. 44(3):881–887.

Sevilla I, Garrido JM, Geijo M, Juste RA. 2007. Pulsed-field gel electrophoresis profile homogeneity of *Mycobacterium avium* subsp. *paratuberculosis* isolates from cattle and heterogeneity of those from sheep and goats. BMC Microbiol. 7(1):18.

Shankar H, et al. 2010. Presence, characterization, and genotype profiles of *Mycobacterium avium* subspecies *paratuberculosis* from unpasteurized individual and pooled milk, commercial pasteurized milk, and milk products in India by culture, PCR, and PCR-REA methods. Int J Infect Dis. 14(2):e121–e126.

Sharma CM, et al. 2010. The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature 464(7286):250–255.

Shine J, Dalgarno L. 1974. The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc Natl Acad Sci U S A. 71(4):1342–1346.

Simpson JT, et al. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19(6):1117–1123.

Sohal J, Singh S, Singh P, Singh A. 2010. On the evolution of 'Indian Bison type' strains of *Mycobacterium avium* subspecies *paratuberculosis*. Microbiol Res. 165(2):163–171.

Sreevatsan S, et al. 1997. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci U S A. 94(18):9869–9874.

Stamatakis A. 2014. RAxMl version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.

Stevenson K, et al. 2002. Molecular characterization of pigmented and nonpigmented isolates of *Mycobacterium avium* subsp. *paratuberculosis*. J Clin Microbiol. 40(5):1798–1804.

Stief B, et al. 2012. Paratuberculosis in a miniature donkey (*Equus asinus f. asinus*). Berl Münch Tierärztl Wochenschr. 7(1–2):38–44.

Stucki D, Gagneux S. 2013. Single nucleotide polymorphisms in *Mycobacterium tuberculosis* and the need for a curated database. Tuberculosis 93(1):30–39.

Thibault VC, et al. 2007. New variable-number tandem-repeat markers for typing *Mycobacterium avium* subsp. *paratuberculosis* and *M. avium* strains: comparison with IS900 and IS1245 restriction fragment length polymorphism typing. J Clin Microbiol. 45(8):2404–2410.

Thorel MF, Krichevsky M, Lévy-Frébault VV. 1990. Numerical taxonomy of mycobactin-dependent mycobacteria, emended description of *Mycobacterium avium*, and description of *Mycobacterium avium* subsp. *avium* subsp. nov., *Mycobacterium avium* subsp. *paratuberculosis* subsp. nov., and *Mycobacterium avium* subsp. *silvaticum* subsp. nov. Int J Syst Bacteriol. 40(3):254–260.

Tian C, Jian-Ping X. 2010. Roles of PE_PGRS family in *Mycobacterium tuberculosis* pathogenesis and novel measures against tuberculosis. Microb Pathog. 49(6):311–314.

Trotochaud AE, Wassarman KM. 2004. 6S RNA function enhances long-term cell survival. J Bacteriol. 186(15):4978–4985.

Trotochaud AE, Wassarman KM. 2006. 6S RNA regulation of pspF transcription leads to altered cell survival at high pH. J Bacteriol. 188(11):3936–3943.

Turenne CY, Collins DM, Alexander DC, Behr MA. 2008. *Mycobacterium avium* subsp. *paratuberculosis* and *M. avium* subsp. *avium* are independently evolved pathogenic clones of a much broader group of *M. avium* organisms. J Bacteriol. 190(7):2479–2487.

Turenne CY, Semret M, Cousins DV, Collins DM, Behr MA. 2006. Sequencing of *hsp65* distinguishes among subsets of the *Mycobacterium avium* complex. J Clin Microbiol. 44(2):433–440.

van Pittius NCG, et al. 2006. Evolution and expansion of the *Mycobacterium tuberculosis* PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. BMC Evol Biol. 6(1):95.

van Soolingen D, Hermans P, De Haas P, Soll D, Van Embden J. 1991. Occurrence and stability of insertion sequences in *Mycobacterium tuberculosis* complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. J Clin Microbiol. 29(11):2578–2586.

Wehner S, Damm K, Hartmann RK, Marz M. 2014. Dissemination of 6S RNA among bacteria. RNA Biol. 11(11):1467–1478.

Wehner S, Mannala GK, et al. 2014. Detection of very long antisense transcripts by whole transcriptome RNA-seq analysis of *Listeria monocytogenes* by semiconductor sequencing technology. PLoS One. 9(10):e108639.

Weinberg Z, et al. 2007. Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. Nucleic Acids Res. 35(14):4809–4819.

Weinberg Z, et al. 2010. Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. Genome Biol. 11(3):R31.

Whittington R, Hope A, Marshall D, Taragel C, Marsh I. 2000. Molecular epidemiology of *Mycobacterium avium* subsp. *paratuberculosis*: IS900 restriction fragment length polymorphism and IS1311 polymorphism analyses of isolates from animals and a human in Australia. J Clin Microbiol. 38(9):3240–3248.

Whittington RJ, Marshall DJ, Nicholls PJ, Marsh IB, Reddacliff LA. 2004. Survival and dormancy of *Mycobacterium avium* subsp. *paratuberculosis* in the environment. Appl Environ Microbiol. 70(5):2989–3004.

Winkler W, Nahvi A, Breaker RR. 2002. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. Nature 419(6910):952–956.

Wu CW, Glasner J, Collins M, Naser S, Talaat AM. 2006. Whole-genome plasticity among *Mycobacterium avium* subspecies: insights from comparative genomic hybridizations. J Bacteriol. 188(2):711–723.

Wynne JW, et al. 2010. Resequencing the *Mycobacterium avium* subsp. *paratuberculosis* k10 genome: improved annotation and revised genome sequence. J Bacteriol. 192(23):6319–6320.

Wynne JW, et al. 2011. Exploring the zoonotic potential of *Mycobacterium avium* subspecies *paratuberculosis* through comparative genomics. PLoS One 6(7):e22171.

Yakrus MA, Good RC. 1990. Geographic distribution, frequency, and specimen source of *Mycobacterium avium* complex serotypes isolated from patients with acquired immunodeficiency syndrome. J Clin Microbiol. 28(5):926–929.

Yusuf D, Marz M, Stadler P, Hofacker I. 2010. Bcheck: a wrapper tool for detecting RNase P RNA genes. BMC Genomics 11(1):432.

Zerbino DR, Birney E. 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. 18(5):821–829.

Associate editor: Tal Dagan